句法特征在动词词义排歧中的应用[1]

王宏显　周强

清华大学信息技术研究院语音和语言技术中心

清华信息科学技术国家实验室技术创新与开发部语音和语言技术中心

清华大学计算机科学与技术系，北京，100084

Email: wanghongxian@gmail.com, zq-lxd@tsinghua.edu.cn

摘要：特征选取是用统计方法进行词义排歧的关键。本文通过句法块的分析结果，将主语和宾语、以及主语和宾语在知网中的归类信息应用于目标动词的词义排歧。实验表明，句法特征对于目标动词的词义排歧有重要作用，相对于仅使用词语和词性特征的系统，加入人工标注句法特征后，拥有类动词正确率由 83.7%提高到 89.2%，存在类动词正确率由 84.8%提高到 86.1%。

关键词：句法特征，词义排歧，最大熵，主语，宾语

Using Syntactic Features in Verb Sense Disambiguation

Wang Hongxian　Zhou Qiang

Center for Speech and Language Technologies, Research Institute of Information Technology, Tsinghua University

Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Information Science and Technology

Department of Computer Science and Technology, Tsinhgua University, Beijing 100084, China

Email: wanghongxian@gmail.com, zq-lxd@tsinghua.edu.cn

Abstract: Feature selection is the key step in statistical method based word sense disambiguation. Based on the result of chunk parsing, we apply the subject, object and their categories in HowNet to target verb sense disambiguation. The experimental result shows that, the syntactic feature is useful for target verb sense disambiguation. Compared with using word and POS feature only, the accuracy increases from 83.7% to 89.2% after the syntactic feature added.

Keywords: Syntactic feature, Word Sense Disambiguation, Maximum Entropy

1 引言

歧义性与模糊性是自然语言的一大特点。一个词语，在不同的句子中可以有不同的语义。这取决于其搭配词语，句法结构，甚至是篇章的主题。比如，“这是一个红色政权。”和“这是一个红色房子。”在两个句子中，不同之处只是替换了一个词语，然而红色的含义却是不同的。修饰政权的是说政权是共产主义性质的，用的是引申义；而修饰房子则是说房子的颜色是红色，用的是其本义。在这里，我们可以根据其修饰的对象判断其义项。但是有些时候，我们只能通过更多的上下文才能判断。比如说，“这是一本红色杂志”。可能是说杂志封面是红色的，也可能是说杂志内容在宣扬共产主义思想。我们需要更多的信息才能判断。

由于自然语言的歧义性，给计算机理解人类语言造成了很大困难。在信息抽取(Stokoe and Oakes et al., 2003)、文本摘要、机器翻译(Chan and Ng et al., 2007)等领域都会遇到这个问题。

[1] 本项研究得到了国家自然科学基金（项目号：60573185, 60873173），国家 863 高科技发展计划（项目号：2007AA01Z173）和清华-Intel 合作研究项目经费的资助。

而词义排歧正是要解决这一问题，为真实文本中的词语确定一个义项。针对这一问题，前人已经做过大量的工作。采用的方法不外乎自然语言处理领域的两大方法，基于规则和基于统计。基于规则的方法由于其规则编写工作量大，自然语言复杂多变，难以全面涵盖，可扩展性差等弊端，已渐渐淡出研究者的视野。而统计方法由于其在这些方面的优点，成为了研究者的主要关注点。

统计方法主要有三个要素：统计学习模型，训练语料和特征选取策略。特征选择是统计学习的一个重要方面，也是应用领域学者主要关注的方面。特征选择可以细分为特征选择与特征表示两个方面。同样的特征，不同的表示方法所包含的信息量不同，好的表示方法可以表达特征中隐含的信息，而不好的方法可能丢失某些信息。

具体到词义排歧，根据汉语的特点，我们可以从字、词、句、篇章等几个方面提取所需的特征。这几种特征的提取难度依次提高的，在以往的研究中，主要是从词语以及词性的角度提取特征，在句法角度研究较少。

本文所做的是句子中目标动词词义排歧，因而难以提取到篇章方面的信息。在后文中，我们在词语和词性特征的基础上，通过对比加入句法特征前后的排歧效果，说明句法特征在目标动词词义排歧中的效果。

2 相关工作

自从 1991 年 Brown(Brown and Pietra et al., 1991)把统计模型引入词义排歧研究以来，研究者已经对基于统计方法的词义排歧进行了广泛的研究。决策树(Decision Tree)、支持向量机(Support VectorMachine, SVM) 、最大熵(Maximum Entropy, ME)等都是常用的模型。

Senseval作为词义排歧方面的评测，为词义排歧技术的发展提供了展示和交流的平台。由Senseval发展而来的SemEval同样关注于词义处理。在SemEval2007中首次加入了中文词义排歧的任务(SemEval-2007 task 5 Multilingual Chinese-English Lexical Sample Task)(Jin and Wu et al., 2007)。在这次评测任务中，参赛者在特征的选取上较为全面，但主要集中在局部特征中的词语和词性标记上。如(Katz and Singleton et al., 2007)使用了目标词周围的一元、二元和三元词语特征，(Kwong, 2007)除使用目标词周围的词语特征之外，还使用了词性特征，并且窗口范围也扩大到了前后10个词语。(Niu and Ji et al., 2007)除目标词语周围的词语特征外，还使用了目标词周围的词语搭配特征，也取得了不错的效果。(Xing, 2007)则使用了目标词周围的词语、词性以及通过浅层句法分析得到句法块特征，取得了这次评测的最好成绩。从评测的结果来看，句法特征对于目标词语的词义排歧是有一定效果的。在此基础上，根据动词的特点，本文着重探索了动词的主语和宾语特征对目标动词词义排歧的效果。

3 语料库与系统描述

3.1 语料库

本文使用清华树库(TCT)中的“存在”和“拥有”两类动词框架标注作为基础语料。该语料库针对动词进行标注，每个句子标出了目标动词、句子中的每个词语词性以及目标动词在知网、同义词词林、现汉通等多种词典中的义项。其中，“存在”类动词标注共有 5519 句，“拥有”类

动词共有 10046 句，在实验中，我们按 2:1 的比例分别将两个文件划分为训练集和测试集。划分按照每个目标动词随机划分。“拥有”类共有多义动词 53 个，平均义项数 2.1 个。“存在”类有多义动词 89 个，平均义项数 2.4 个。在本文中，我们选用知网义项作为义项表示。另外，该语料库还标注了目标动词驱动的句法块，包括目标动词的施事、受事、状语等。这为我们研究句法特征对于词义排歧的作用提供了基础。一个典型语义标注结构如下(省略了与本文无关的信息)。

目标动词= 包容

动词位置= 6

基本标注= 这/rN 种/qN 不同/a 语种/n 中/f 词义/n 包容/vN 范围/n 的/u 差异/n ，/， 有时/d 引起/v 国际/n 学术/n 交流/vN 中/f 的/u 周折/n 。/。

知网义项描述= 4869- {contain|包含}

词林义项描述= Jd050101-包容,撑肠拄肚,兼容并包,容,容纳,容受,盛

现汉通义项描述= 容纳。

清华库义项描述= 容纳

句法语义块标注= 这/rN 种/qN 不同/a 语种/n 中/f [A-np-DZ 词义/n-@ 包容/vN-@]Tgt [H-np-SG 范围/n-@]y 的/u 差异/n ，/， 有时/d 引起/v 国际/n 学术/n 交流/vN 中/f 的/u 周折/n 。/。

其中-@指出了各语义块的中心词。

3.2 系统描述

本文的处理采用最大熵模型。最大熵模型无需考虑各个特征之间的独立性，多个相关或不相关的特征可以同时使用，这有利于使用多种特征增强分类器的性能。最大熵模型可以直接使用文本特征，每个特征以一个字符串出现。用空格隔开的多个字符串组成了特征向量。本文采用的是 ZhangLe 的最大熵工具包，针对每个词语训练单独的分类器。进行排歧时，根据目标动词选择相应的分类器，给出目标动词的义项。训练中使用的高斯平滑，模型中的平滑参数为 1。

与多数基于统计的词义排歧系统类似，本文采用的基本特征为目标动词、目标动词的前后各两个词语、目标动词前后各两个词语的词性。其抽取模板如表 1，我们使用的全部是 unigram 特征。

词语特征	W-2， W-1， W0， W1， W2
词性特征	P-2， P-1， P0， P1， P2

表 1 基本特征模板

3.3 评价方法

参照 SemEval2007 Task5 中的评价方法，我们使用 MicroAve(Micro-Average Accuracy)和 MacroAve(Macro-Average Accuracy)两个参数对结构进行评价。这两个参数通过下面的公式定义：

$$MicroAve = \frac{\sum_{i=1}^{N} m_i}{\sum_{i=1}^{N} n_i} \quad (1)$$

$$MacroAve = \frac{\sum_{i=1}^{N} m_i / n_i}{N} \quad (2)$$

其中，N 表示总词数，m_i表示第 i 个词排歧正确的词语，n_i表示第 i 个次出现的总次数。

4 句法特征选择

动词有一些自身的特点，一般都有主语和宾语，或者存在其中之一。动词的词义与其所带的主语和宾语密切相关，主语和宾语发生变化，动词的词义往往也会发生变化。也就是说，动词的词义往往是由其所处的句法结构，以及其所支配的结构框架中的词语决定的。基于这样的考虑，我们在上一节所述的基本特征之外，提取动词所在句法结构中的词语作为词义排歧的重要特征。提取的特征主要包括主语（宾语）是否存在，主语（宾语）中心词以及其上位义原和所属类别等。实验中使用的句法分析器是宇航(宇航, 2007)开发的句法块分析器。该句法块分析器标出了义目标动词为中心的句法框架信息，并对每个成分标出了中心语。其在 TCT 树库上的 F 值为 85%。标注举例如下：

目标动词= 逼

动词位置= 3

事件标注= 兼并/v 联合/v ：/wM [P-vp 逼/v-@ 出来/vB] 的/uJDE [H-np 思路/n-@]

说明：@表示句法块中心词，短杠前的 P、H 为动词框架角色标记，短杠后的 vp、vp 等为句法功能标记。

从上述标注中，我们可以提取到目标动词的框架结构信息，主要是主语和宾语相关的信息。具体来说，提取的特征有以下八种。

1) 主语是否存在（subject-exist）；
2) 主语中心词（subject-word）；
3) 主语是否为一个短语结构(subject-IsPhrase)；
4) 宾语是否存在（object-exist）；
5) 宾语中心词（object-word）；
6) 宾语是否为一个短语结构(object-IsPhrase)；
7) 主语和宾语中心词的上位义原(subject-up, object-up)；
8) 主语和宾语中心词类别，有两个独立划分标准（具体名词/抽象名词，人/物）。

以上特征并非全部同时存在，很多情况下只能提取其中的部分特征。例如，主语不存在时，无法提取主语中心词。遇到无法提取特征的情况，我们的策略是简单放弃这一特征。特征 3 和 6 中，判断主语或宾语是否为短语结构的标准是，考查主语(宾语)块是否由多个词语组成，如果多于一个词，认为是短语结构；如果只有一个词，则认为是非短线与结构。特征 7 中的上位义原通过查询知网义项描述得到。特征 8 中的中心语类别的划分也是以知网为标准。如果一个词语通过上位义原不断上溯至{entity|实体}的过程中经过了{physical|物质}认为其是具体名词，否则为抽象名词。划分人与物的标准类似，如果一个词语在上溯过程中经过了{human|人}，认为该词语是“人”名词，否则认为是“物”名词。

5 实验结果分析

实验以只使用词语和词性特征的系统作为 BaseLine，然后依次使用如下三种特征，并与 BaseLine 比较。

1) 词语特征，词性特征，自动标注句法特征；
2) 词语特征，词性特征，自动标注句法特征，知网特征；
3) 词语特征，词性特征，人工标注句法特征，知网特征。

其中，句法特征是指上一节中提到的特征 1-6，知网特征指上节中的特征 7-8。实验结果如表 2 所示。

处理方法	Exist Micro-Ave	Exist Macro-Ave	Have Micro-Ave	Have Macro-Ave
词语和 POS	0.848	0.756	0.837	0.766
加入句法结构	0.857	0.774	0.859	0.764
加入 HowNet	0.858	0.774	0.867	0.775
使用人工标注	0.861	0.781	0.892	0.808

说明：exist 指存在类标注文件，have 指拥有类标注文件。

表 2 高频词不同特征下正确率比较

我们看到，每次加入新的特征之后，两个标注文件的正确率都有所提升。在使用自动句法块标注的条件下，存在类动词正确率由 84.8%提高到 85.8%，拥有类动词正确率由 83.7%提高到 86.7%。考虑到自动句法标注会有一定的噪声，在实用人工句法块标注的条件下，正确率会有进一步的提升。这证实了我们根据之前实验所进行的推论。从句法结构中所提取的特征对动词的词义排歧有较大的正面作用。

从整体上看，拥有类标注文件正确率增加的更多。我们注意到，使用人工句法标注后，拥有类的正确率有显著提升（提高 3%）。通过分析两个标注文件中动词的差异，对这个现象的一个合理解释是，拥有类动词一般都同时有主语和宾语，而且其中的主语往往是“人”名词。这样，经过句法分析之后，这些对动词词义的确定有重要作用的主语和宾语特征都被提取出来，从而提高了排歧正确率。而使用人工标注句法结构之后正确率有显著提高，正说明了动词词义对其主语或宾语的敏感性。

下面我们通过对正确率有变化具体词语进行分析，说明句法结构特征在动词词义排歧中的作用。由于训练词例较少的词语随机误差较大，我们从拥有类标注文件选取训练词例较多、训练比较充分的进行分析。表 3 是拥有类标注文件中词频最高的前 9 个词语。

这些词语中绝大部分都有一定程度的提高，这表明了句法信息对词义排歧的有效性。我们通过分析目标动词“找”在仅使用词语和词性特征时排歧错误，使用人工句法标注和知网特征后排歧正确的例子，说明句法特征在动词词义排歧中的作用。

“找”的两个义项{LookFor|寻}和{request|要求:ResultEvent={meet|会面}}，主要通过宾语是否是人来区分。通过句法分析后，原来的大部分错误都排除了，在自动句法标注条件下，正确率提高了 11.6%，而在人工标注条件下正确率提高了 18.9%。下面是在仅使用词语和词性特征的条件下排歧错误的例子。

句子= 他 为 自己 历经 忧患 ， 艰难 探索 而 找 到 真理 感到 无比 激动 和 喜悦 。
动词= 找
人工标注义项= {LookFor|寻}
义项= {request|要求: ResultEvent={meet|会面}}

词语	义项数	词数	词语+POS	加入句法信息	加入 HowNet	人工句法标注	正确率提高
有	2	1123	0.813	0.847	0.862	0.890	0.077
发展	2	530	0.996	0.996	0.996	0.996	0
找	2	95	0.695	0.811	0.8	0.884	0.189
感到	2	71	0.690	0.718	0.704	0.761	0.071
给	3	62	0.613	0.677	0.661	0.710	0.097
开发	2	41	0.854	0.902	0.854	0.927	0.073
属于	2	36	0.750	0.722	0.722	0.722	-0.028
处理	2	32	0.875	0.875	0.875	0.875	0
转	3	28	0.571	0.643	0.643	0.643	0.072

表 3 拥有类标注文件中词频较大的词语

经过句法分析后，确定了“真理”为宾语的中心，并且是一个物体名词。从而确定其词义为{LookFor|寻}。这表明句法分析确定了目标动词的论元，对目标动词词义的确定有很大的帮助。

6 结论与展望

. 特征选取是基于统计模型的词义排歧系统的关键环节，在以往的工作中，或因为语料库的限制，或因为句法分析工具的限制，对句法特征在词义排歧中的作用研究较少。本文通过从自动句法标注结果和人工句法标注结果中抽取目标动词所支配的主语和宾语中心词，并依据知网对其进行归类，以此作为句法特征应用目标动词的词义排歧，取得了较好的实验效果。证明句法特征对于动词词义排歧是十分有效的。我们所使用的句法分析器只需分析出基本的块结构，具有很高的效率，这也是句法特征可以应用于词义的基础。

由于句法库的有限性，我们只针对有限的两类语义范畴的动词进行了实验，在后续的工作中，我们将考虑将实验范围扩大到所有动词以及其他类型词语，考察句法特征对词义排歧的效果。

参考文献

Brown, P. F. and S. A. D. Pietra, et al. (1991). Word-sense disambiguation using statistical methods. Proceedings of the 29th annual meeting on Association for Computational Linguistics , Berkeley, California , Association for Computational Linguistics.

Chan, Y. S. and H. T. Ng, et al. (2007). Word sense disambiguation improves statistical machine translation. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics(ACL).

Jin, P. and Y. Wu, et al. (2007). SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample.

Katz, P. and M. Singleton, et al. (2007). SWAT-MP: The SemEval-2007 Systems for Task 5 and Task 14.

SemEval2007.
Kwong, O. Y. (2007). CITYU-HIF: WSD with Human-Informed Feature Preference. SemEval2007.
Niu, Z. and D. Ji, et al. (2007). I2R: Three Systems for Word Sense Discrimination, Chinese Word Sense Disambiguation, and English Word Sense Disambiguation. SemEval-2007.
Stokoe, C. and M. P. Oakes, et al. (2003). Word sense disambiguation in information retrieval revisited. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval , Toronto, Canada , ACM.
Xing, Y. (2007). SRCB-WSD: Supervised Chinese Word Sense Disambiguation with Key Features. SemEval2007.
宇航（2007）．汉语句法块自动分析研究，清华大学本科毕业设计论文．2007年6月．